令和6京都大学化学研究所　スーパーコンピュータシステム　利用報告書

Mapping species boundaries among the Nucleocytoviricota using high-resolution phylogenomics

高解像度系統ゲノム解析を用いた核細胞ウイルス門の種境界のマッピング

京都大学化学研究所 緒方博之研究室
NECHES Russell Young（ネチェス ラッセル ヤン）

研究成果概要

Genomes, MAGs and contigs were selected from the high quality fraction of the IMG/VR 4.1 database, consisting of 5,576,197 uncultivated viral genomes (UViGs) in 5,621,398 contigs. Records assigned to the phylum Nucleocytoviricota by the IMG/VR analytical pipeline were selected, yielding 104,220 UViGs in 149,421 contigs. ANI scores were computed between all pairs of genomes using `fastANI` 1.34 with a kmer size of 16, a fragment length of 3000, a minimum of five bidirectional fragment mappings, and a minimum aligned fraction of 20%. ANI clusters were built using single linkage hierarchical clustering using the scipy 1.13.0 at clustering thresholds from 77% to 99% in increments of 1%. Open reading frames within these contigs were predicted using prodigal-gv 2.11.0, yielding 3,028,373 predicted proteins. Ortholog groups were identified by searching these proteins with the 8863 hidden Markov models (HMMs) of giant virus orthologous group (GVOG) database using hmmer 3.4, yielding 12,968,467 hits with an e-value less than 10^{-6}.

When a genome is hit by the same GVOG HMM more than once, the paralogous genes were ranked in order of their e-values. When more than one HMM aligned to the same predicted protein sequence, the sequence was assigned to GVOG whose model produced the lower e-value, with the higher bitscore as a tie-breaker. For the genomes belonging to each ANI cluster, the nucleotide sequences of each gene belonging to each ortholog group were aligned using MAFFT 7.526 and trimmed using trimAl 1.4.1, which selects heuristics optimized for maximum likelihood phylogenetic tree reconstruction. In this mode, sequences that lack informative sites are removed. Sequences with an aligned fraction of less than 75% were removed. Alignments with genes from fewer than five genomes after trimming and filtering were dropped. For each ortholog group, for each genome ANI cluster, for each genome ANI clustering cutoff threshold, a phylogenetic tree was inferred using fasttree 2.1.11. For clusters in which at least five genomes shared at least two ortholog groups, the phylogenetic structures of every pair of ortholog group was compared using SuchTree 1.0. Because the IMG/VR database is synthesized from a large number of metagenomic assembled genomes (MAGs) curated from many different sources, only orthologs observed on the same contig were treated as linked for the purposes of phylogenomic analysis to mitigate the potential impact of bin contamination. When paralogs were present, only the paralog that scored the lowest e-value by its associated GVOG HMM was considered. The analytical workflow was implemented as a reproducible Snakemake workflow, available on GitHub under the name kizuchi.